Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention

L. Itti,[1] N. Dhavale[1] and F. Pighin[2]

[1] Department of Computer Science, University of Southern California, Los Angeles, California
[2] Intitute for Creative Technologies, University of Southern California, Los Angeles, California

ABSTRACT

We describe a neurobiological model of visual attention and eye/head movements in primates, and its application to the automatic animation of a realistic virtual human head watching an unconstrained variety of visual inputs. The bottom-up (image-based) attention model is based on the known neurophysiology of visual processing along the occipito-parietal pathway of the primate brain, while the eye/head movement model is derived from recordings in freely behaving Rhesus monkeys. The system is successful at autonomously saccading towards and tracking salient targets in a variety of video clips, including synthetic stimuli, real outdoors scenes and gaming console outputs. The resulting virtual human eye/head animation yields realistic rendering of the simulation results, both suggesting applicability of this approach to avatar animation and reinforcing the plausibility of the neural model.

1. INTRODUCTION

Animating realistically the human face is one of computer graphics' greatest challenges. Realistic facial animation has spurred a great wealth of research, from motion capture to hair animation. Most of the research in facial animation has focused on speech,[1–3] while the motions of some of the face components have been neglected, in particular the eyes and the rigid motion of the head. These two motions combined define the gaze behavior of the face. Animals, however, have developed an impressive ability to estimate where another animal's gaze is pointed to. Humans, for instance, can determine gaze direction with an accuracy better than 4°.[4] Furthermore, gaze estimation is an essential component of non-verbal communication and social interaction, to the point that monitoring another's eye movements is often believed to provide a window onto that other person's mental state and intentions.[5] This crucial role of gaze estimation in determining intentionality and in interacting with social partners makes any inaccuracy in its rendering obvious to even the casual observer.

Animating an avatar's gaze is a challenging task. The human eye and its control systems indeed are extremely complex.[6,7] Contrary to conventional video cameras which uniformly sample the visual world, the human eye only gathers detailed visual information in a small central region of the visual field, the fovea. Increasingly peripheral locations are sampled at increasingly coarser spatial resolutions, with a complex-logarithmic fall-off of resolution with eccentricity.[8] This over-representation of the central visual field, with the fovea only subtending $2-5^\circ$ of visual angle, may be understood through considerations of information processing complexity. Indeed, if the entire $160 \times 175^\circ$ subtended by each retina[9] were populated by photoreceptors at foveal density, the optic nerve would comprise on the order of one billion nerve fibers, compared to about one million fibers in humans.[9] Rather than growing the optic nerve and subsequent cortical visual processing area 1,000-fold or more, nature has devised an alternate strategy to coping with visual information overload: The under-representation of the visual periphery is compensated by the presence of highly mobile eyes that can almost instantaneously point the fovea towards any location in the visual world. Given this architecture, developing efficient control strategies for the deployment of gaze onto behaviorally-relevant visual targets has become a problem of pivotal importance to the survival of the organism. These complex strategies are very difficult to simulate accurately.

Recording eye motion is also a difficult task. Optical motion capture systems traditionally used to record facial motions fail to robustly track eye movements. Atempts have been made through the use of contact lenses coated with retroreflective material. However, lenses have a tendency to move on the surface of the eye, thus preventing from recording accurate motions. Systems specific for recording eye movements typically can accurately track eye motions, in a controlled environment. Unfortunately they tend to be expensive ($\$20,000$ for an Iscan eye tracker) and intrusive, which precludes their use in a wide range of situations. For these reasons we are following a procedural approach. Our animation technique heavily relies on research performed in the field of neuroscience.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**15**	

In our approach we thus leverage studies conducted in neuroscience to build a procedural gaze animation system driven exogenously by external visual stimuli. How primates, including humans, select where to look next in the presence of various visual stimuli indeed has been one of the most active topics of neuroscience research over the past century. Key concepts have emerged, including that of focal visual attention. Attention complements explicit (so-called "overt") gaze shifs with a rapidly shiftable "virtual spotlight"[10] that covertly selects stimuli in turn for detailed processing and access to conscious representation. Thus, understanding how gaze is directed requires first an understanding of how attention may be attracted towards specific visual targets. While introspection easily reveals that attention may be guided voluntarily – or from the "top-down" – (e.g., attending to that attractive mysterious stranger while talking to this tired old friend at a cocktail party), its deployment has also been shown to be strongly influenced from the "bottom-up," by the very contents of the visual input.[11–13] For instance, an object thrown at us, or the abrupt onset of a bright light, will often capture attention and elicit an almost reflex-like eye movement towards that object or light. In normal vision, both the very fast (able to scan up to 20 items/s) bottom-up and much slower (limited to about 5 items/s) top-down modes of attentional guidance cooperate, in an attempt to efficiently analyze the incoming visual information in a manner that is both serving behavioral priorities and keeping the organism alert of possible unexpected dangers.

Here we explore the applicability of a neurobiological model of visual attention to the automatic realistic generation of eye and head movements given arbitrary video scenes. Technological applications of this model include interactive games, character animation in production movies, human-computer interaction and others. Because the neural basis of top-down attention remains elusive, our model primarily implements a bottom-up guidance of attention towards conspicuous visual targets, augmented by a realistic eye/head movement controller and an eyeblink generator. We use this model to animate a photorealistic 3D model of a human face. The system is tested with arbitrary video clips and autonomously predicts the motion of the eyes, head, eyelids, and a number of accompanying deformations of the virtual facial tissue.

2. RELATED WORK

2.1. Modeling Visual Attention

Under the broad umbrella of attention, several key aspects of sensory processing have been defined, such as the anatomical and functional seggregation in the monkey brain between localization of salient targets ("where/how" dorsal pathway) and their recognition ("what" ventral pathway)[14]). The very existence of this segregation has made it a reasonable enterprise to study attention separately from object recognition. There are many other aspects of attention, which we will not consider here, such as its role as an information processing bottleneck,[15] its role in binding the different visual attributes of an object, such as color and form, into a unitary percept,[11] and its spatially- and feature-selective modulatory effect on low-level visual processing[16,17] (we refer the reader to recent reviews for further information[13,18]).

Development of computational models of attention started with the Feature Integration Theory of Treisman *et al.*,[11] which proposed that only fairly simple visual features are computed in a massively parallel manner over the entire incoming visual scene. Attention is then necessary to bind those early features into a united object representation, and the selected bound representation is the only part of the visual world that passes though the attentional bottleneck.

The first explicit neural architecture for the bottom-up guidance of attention was proposed by Koch and Ullman,[12] and is closely related to the feature integration theory. Their model is centered around a saliency map, that is, an explicit two-dimensional topographic map that encodes for stimulus conspicuity, or salience, at every location in the visual scene. The saliency map receives inputs from early visual processing, and provides an efficient centralized control strategy by which the focus of attention simply scans the saliency map in order of decreasing saliency.

This general architecture has been further developed by Itti *et al.*[19,20] to yield a full computer implementation, used as the basis for our work. In this model, the early stages of visual processing decompose the visual input through a set of feature-selective filtering processes endowed with contextual modulatory effects. To control a single attentional focus based on this multiplicity of visual representation, feature maps are combined into a unique scalar saliency map. Biasing attention towards the location of highest salienceis then achieved by a winner-take-all neural network, which implements a neurally distributed maximum detector. Once attended to, the current location is transiently inhibited in the saliency map by an inhibition-of-return mechanism. Thus, the winner-take-all naturally converges to the next most salient location, and repeating this process generates attentional scanpaths.

Many other models have been proposed, which typically share some of the components of that just described; in particular, a saliency map is central to most of these models[21–23] (see[13] for a more in-depth overview). In the present work,

we extend the model of Itti *et al.* (available online in C++ source-code form) from the prediction of covert attentional shifts onto static scenes to that of eye and head movements onto dynamic video clips.

2.2. Modeling Eye and Head Movements

Behavioural studies of alert behaving monkeys and humans have indicated that gaze shifts are accomplished by moving both the eyes and head in the same direction.[24,25] Sparks[7] provides a comprehensive review on the topic. Of particular interest here, Freedman and Sparks[26] specifically investigated saccadic eye movements made by *Rhesus* monkeys when their head was unrestrained. Remarkably, they found that the relative contribution of the head and eyes towards a given gaze shift follows simple laws. Their model is at the basis of the eye/head saccade subsystem implemented in our model, and is described in details in a further section.

Video analysis under various conditions has revealed that normal eye blinking rate is 11.6/min, while in the attentive state this increases slightly to 13.5/min in normal human subjects, with faster rates of 19.1/min and 19.6/min respectively observed in schizophrenic patients.[27,28]

2.3. Gaze Animation in Avatars

Several approaches have been proposed to endow avatars with realistic-looking eye movements. They primarily fall under three categories, for each of which we describe a representative example system below.

A first approach consists of augmenting virtual agents with eye movements that are essentially random, but whose randomness is carefully matched to the observed distributions of actual eye movements produced by humans. In a recent example of this approach, Lee *et al.*[29] used an eye-tracking device to gather statistical data on the frequency of occurrence and spatiotemporal metric properties of human saccades. Random probability distributions derived from the data were then used to augment the pre-existing behavior of an agent with eye movements. This model makes no attempt to direct eye movements towards specific targets in space, beyond the fact that gaze direction either deterministically follows large head movements or is randomly drawn from the empirically measured distributions. Thus, Lee *et al.*'s remarkably realistic model addresses a fundamentally different issue from that investigated here: While we focus on automatically determining where the avatar should look, Lee *et al.* add very realistic-looking random eye movements to an avatar that is already fully behaving.

A second approach consists of using machine vision to estimate gaze and pose from humans and to use those estimated parameters to drive a virtual agent.[30–32] Various machine vision algorithms are used to extract critical parameters from video acquired from one or more cameras pointed to a human. These parameters are then transmitted to the rendering site and used for avatar animation. By definition, this type of approach is only applicable to situations where the avatars mimic or embody humans, such as video conferencing, spatially-distributed online gaming and human-computer interfacing. This substantially differs from our goal here, which is to autonomously animate agents.

A last approach consists of using machine vision to attempt to locate targets of interest in virtual or real scenes.[33–35] For instance, Terzopoulos and Rabie[36] have proposed an active vision system for animats behaving in virtual environments, including artificial fish and humans. The animats are equipped with multiscale virtual cameras, providing both foveal and wide-angle visual input. Selection of targets for eye movements is based on a color signature matching algorithm, in which objects known from their color histogram are localized in the image through backprojection. A set of behavioral routines evaluate sensory inputs against current mental state and behavioral goals, and decide on a course of motor action, possibly including saccadic eye movements to detected objects. Temporal dynamics and sharing between head and eye movements are not described in details. This system presents limitations, in particular in its sensory abilities (e.g., it will not saccade to novel objects of unknown color signature, nor prioritize at the sensory level among multiple objects). While this approach is thus dedicated to a set of known objects with distinct color signatures present in a relatively uncluttered environment such as deep oceans, in its gross architecture it is particularly relevant to the model presented here.

Our approach directly follows the third category described here, using machine vision to determine where to look next. Our more detailed biologically-inspired modeling of the attention and eye movement subsystems allow us to extend this basic paradigm to a wide repertoire of combined eye/head movements in unconstrained environments, containing arbitrary numbers of known or unknown targets against arbitrary amounts of known or unknown clutter.

3. NEUROBIOLOGICAL MODEL OF ATTENTION

3.1. Foveation and Retinal Filtering

Our simulations so far have used pre-recorded video clips as visual inputs, either filmed using a video camera, or captured from the video output of gaming consoles. Interlaced video was digitized using a consumer-electronics framegrabber board (WinTV Go, Hauppage Inc.) at a resolution of 640×480 pixels and 30 frames/s.

Although using pre-recorded video may suggest open-loop processing (i.e., eye movements did not influence camera viewpoint), our simulations actually operated in a spatially-limited closed-loop due to the first step of processing: A gaze-contingent and eccentricity-dependent blur aimed at coarsely reproducing the non-uniform distribution of photoreceptors on the human retina. Efficient implementation of the foveation filter was achieved through interpolation across levels of a Gaussian pyramid[37] computed from each input frame. To determine which scale to use at any image location, we computed a 3/4-chamfer distance map,[38] encoding at every pixel for an approximation to the Euclidean distance between that pixel's location and a disc of 5° diameter centered at the model's current eye position. Thus, pixels close to the fovea were interpolated from high-resolution scales in the pyramid, while more eccentric pixels were interpolated from coarser-resolution scales. Therefore, the raw video clip might be considered as a wide-field environmental view, with the foveation filter isolating a portion of that environment for input to the model. This gaze-contingent foveation ensured more realistic simulations, in which small (or high spatial frequency) visual stimuli far from fixation were unlikely to strongly attract attention.

3.2. Bottom-Up Guidance of Attention

We developed a computational model that predicts the spatiotemporal deployment of gaze onto any incoming visual scene **Fig. 1)**. Our implementation started with the model of Itti *et al.*,[19,20] which is freely available. In addition to the foveation filter just described, our extensions to this model include the addition of a flicker feature that detects temporal change (e.g., onset and offset of lights), of a motion feature that detects objects moving in specific directions, and a number of extensions, described below, to use the model with video rather than static inputs. Additional extensions concern the generation of eye and head movements from the covert attention output of the model, and are detailed in the following section. In what follows, we briefly describe the extended implementation used in our simulations.

3.2.1. Low-Level Feature Extraction

Given a foveated input frame, the first processing step consists of decomposing it into a set of distinct "channels," using linear filters tuned to specific stimulus dimensions. The response properties of the filters were chosen according to what is known of their neuronal equivalents in primates. This decomposition is performed at nine spatial scales, to allow the model to represent smaller and larger objects in separate subdivisions of the channels. The scales are created using Gaussian pyramids[37] with 9 levels (1:1 to 1:256 scaling).

With r_n, g_n and b_n being the red, green and blue channels of input frame n, an intensity image I_n is obtained as $I_n = (r_n + g_n + b_n)/3$. I_n is used to create a Gaussian pyramid $I_n(\sigma)$, where $\sigma \in [0..8]$ is the scale. The r_n, g_n and b_n channels are subsequently normalized by I_n to decouple hue from intensity. Four broadly-tuned color channels are then created: $R_n = r_n - (g_n + b_n)/2$ for red, $G_n = g_n - (r_n + b_n)/2$ for green, $B_n = b_n - (r_n + g_n)/2$ for blue, and $Y_n = r_n + g_n - 2(|r_n - g_n| + b_n)$ for yellow (negative values are set to zero). By construction, each color channel yields maximal response to the pure hue to which it is tuned, and zero response both to black and to white. Four Gaussian pyramids $R_n(\sigma), G_n(\sigma), B_n(\sigma)$ and $Y_n(\sigma)$ are then created.

Local orientation information is obtained from I using oriented Gabor pyramids $O_n(\sigma, \theta)$, where $\sigma \in [0..8]$ represents the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation. (Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope, approximate the receptive field sensitivity profile of orientation-selective neurons in primary visual cortex.[39]) The fast implementation proposed by Greenspan *et al.* is used in our model. [40]

Flicker is computed from the absolute difference between the luminance I_n of the current frame and that I_{n-1} of the previous frame, yielding a flicker pyramid $F_n(\sigma)$.

Finally, motion is computed from spatially-shifted differences between Gabor pyramids from the current and previous frames.[41] The same four Gabor orientations as in the orientation channel are used, and only shifts of one pixel orthogonal to the Gabor orientation are considered, yielding one shifted pyramid $S_n(\sigma, \theta)$ for each Gabor pyramid $O_n(\sigma, \theta)$. Because of the pyramidal representation, this is captures a wide range of object velocities (since a 1-pixel shift at scale 8 corresponds

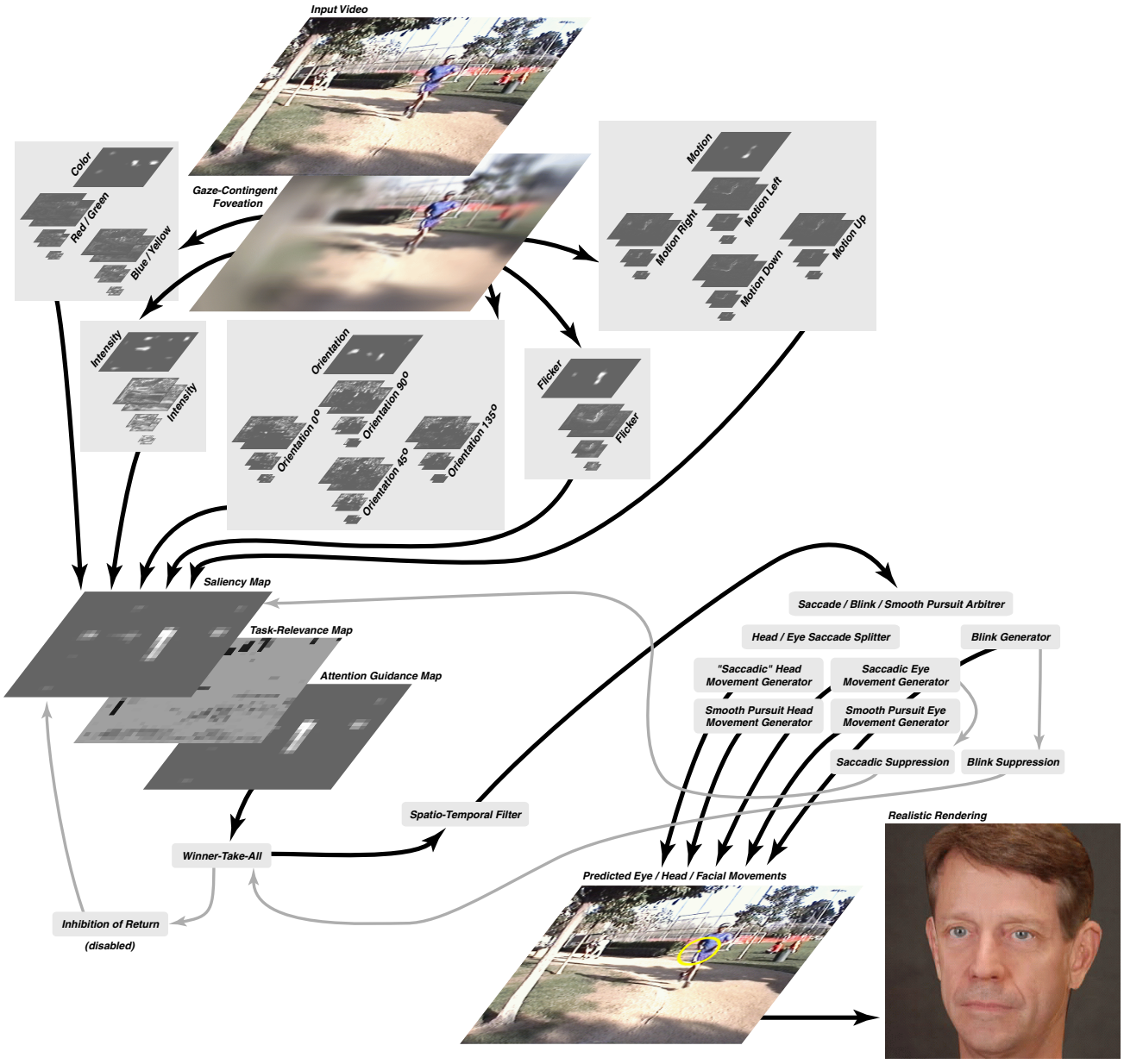

Figure 1. Overview of the model. Input video is processed by a foveation filter followed by low-level feature extraction mechanisms. All resulting feature maps contribute to the unique saliency map, whose maximum is where the model's attention is pointed to. The resulting covert attention shifts drive the eye/head movement controller and realistic facial animation.

to a 256-pixel shift at base scale 0). The Reichardt model for motion computation in the fly brain [42] is then used to compute a motion pyramid $R_n(\sigma,\theta)$:

$$R_n(\sigma,\theta) = |O_n(\sigma,\theta) * S_{n-1}(\sigma,\theta) - O_{n-1}(\sigma,\theta) * S_n(\sigma,\theta)| \quad (1)$$

where * denotes a pointwise product. Values smaller than 3.0 are set to zero to reduce output noise. A fast implementation (used here) uses the intensity rather than orientation pyramid as input (at the cost of losing selectivity for orientation).

Overall, these filters are very crude and yield many false detections, not unlike biological neurons. As we will see below, however, the center-surround and competition for salience applied to their outputs will discard these artifacts.

3.2.2. Center-Surround Receptive Field Profiles

Each feature is computed in a center-surround structure akin to visual receptive fields. This renders the system sensitive to local feature contrast rather than raw feature amplitude.

Center-surround operations are implemented as differences between a fine and a coarse scale for a given feature: The center of the receptive field corresponds to a pixel at level $c \in \{2,3,4\}$ in the pyramid, and the surround to the corresponding pixel at level $s = c + \delta$, with $\delta \in \{3,4\}$. We hence compute six feature maps for each type of feature (at scales 2-5, 2-6, 3-6, 3-7, 4-7, 4-8). Across-scale map subtraction, denoted "$\ominus$," is obtained by interpolation to the finer scale and pointwise subtraction. Nine types of features are computed: on/off intensity contrast,[39] red/green and blue/yellow double-opponency,[43] four orientation contrasts,[44] flicker contrast,[45] and four motion contrasts[42]:

$$\textbf{Intensity:}\quad \mathscr{I}_n(c,s) = |I_n(c) \ominus I_n(s)| \tag{2}$$

$$\textbf{R/G:}\quad \mathscr{RG}_n(c,s) = |(R_n(c) - G_n(c)) \ominus (R_n(s) - G_n(s))| \tag{3}$$

$$\textbf{B/Y:}\quad \mathscr{BY}_n(c,s) = |(B_n(c) - Y_n(c)) \ominus (B_n(s) - Y_n(s))| \tag{4}$$

$$\textbf{Orientation:}\quad \mathscr{O}_n(c,s,\theta) = |O_n(c,\theta) \ominus O_n(s,\theta)| \tag{5}$$

$$\textbf{Flicker:}\quad \mathscr{F}_n(c,s) = |F_n(c) \ominus F_n(s)| \tag{6}$$

$$\textbf{Motion:}\quad \mathscr{R}_n(c,s,\theta) = |R_n(c,\theta) \ominus R_n(s,\theta)| \tag{7}$$

In total, 72 feature maps are computed: Six for intensity, 12 for color, 24 for orientation, 6 for flicker and 24 for motion.

3.2.3. Combining Information Across Multiple Maps

At each spatial location, activity from the 72 feature maps drives a single scalar measure of salience. One major difficulty is to combine *a priori* not comparable stimulus dimensions (e.g., how much should orientation weigh against color?) The crudeness of the low-level filters precludes a simple summation of all feature maps into the saliency map, as the sum would typically be dominated by noise.[20]

To solve this problem, we use a simple within-feature spatial competition scheme, directly inspired by physiological and psychological studies of long-range cortico-cortical connections in V1. These connections, which can span up to 6-8mm, are thought to mediate "non-classical" response modulation by stimuli outside the cell's receptive field, and are made by axonal arbors of excitatory (pyramidal) neurons in cortical layers III and V.[46,47] Non-classical interactions result from a complex balance of excitation and inhibition between neighboring neurons as shown by electrophysiology,[48,49] optical imaging,[50] and human psychophysics.[51]

Our model reproduces three widely observed characteristics of those interactions: First, interactions between a center location and its non-classical surround appear to be dominated by an inhibitory component from the surround to the center,[52] although this effect is dependent on the relative contrast between center and surround.[49] Hence our model focuses on non-classical surround inhibition. Second, inhibition is strongest from neurons tuned to the same stimulus properties as the center.[47,48] As a consequence, our model implements interactions within each individual feature map rather than between maps. Third, inhibition appears strongest at a particular distance from the center,[51] and weakens both with shorter and longer distances. These three remarks suggest that the structure of non-classical interactions can be coarsely modeled by a two-dimensional difference-of-Gaussians (DoG) connection pattern.

Our implementation is as follows: Each feature map $\mathscr{M}$ is first normalized to a fixed dynamic range, in order to eliminate feature-dependent amplitude differences due to different feature extraction mechanisms. $\mathscr{M}$ is then iteratively convolved by a large 2D DoG filter, the original map is added to the result, and negative values are set to zero after each iteration. The DoG filter yields strong local excitation, counteracted by broad inhibition from neighboring locations:

$$\mathscr{M} \leftarrow |\mathscr{M} + \mathscr{M} * \mathscr{DoG} - C_{inh}|_{\geq 0} \tag{8}$$

where $\mathscr{DoG}$ is the 2D Difference of Gaussian filter, $|.|_{\geq 0}$ discards negative values, and C_{inh} is a constant inhibitory term ($C_{inh} = 0.02$ in our implementation with the map initially scaled between 0 and 1). C_{inh} introduces a small bias towards slowly suppressing areas in which the excitation and inhibition balance almost exactly (extended uniform textures). Each feature map is subjected to 10 iterations of the process described in **Eq. 8** and denoted by the operator $\mathscr{N}(.)$ in what follows. This non-linear stage following linear filtering is one of the critical components in ensuring that the saliency map is not dominated by noise even when the input contains very strong variations in individual features.

3.2.4. The Saliency Map

After normalization by $\mathcal{N}(.)$, the feature maps for each feature channel are summed across scales into five separate "conspicuity maps," at the scale ($\sigma = 4$) of the saliency map. They are obtained through across-scale map addition, $\oplus$, which consists of reduction to scale 4 and pointwise addition. For orientation and motion, four intermediary maps are first created combining six feature maps for a given θ, then combined into a single conspicuity map:

$$\textbf{Intensity: } \overline{\mathscr{I}}_n = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathscr{I}_n(c,s)) \tag{9}$$

$$\textbf{Color: } \overline{\mathscr{C}}_n = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathscr{RG}_n(c,s)) + \mathcal{N}(\mathscr{BY}_n(c,s))] \tag{10}$$

$$\textbf{Orientation: } \overline{\mathscr{O}}_n = \sum_{\theta} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathscr{O}_n(c,s,\theta)) \right) \tag{11}$$

$$\textbf{Flicker: } \overline{\mathscr{F}}_n = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathscr{F}_n(c,s)) \tag{12}$$

$$\textbf{Motion: } \overline{\mathscr{R}}_n = \sum_{\theta} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathscr{R}_n(c,s,\theta)) \right) \tag{13}$$

The motivation for the creation of five separate channels and their individual normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map. The conspicuity maps are normalized and summed into the final input $\mathscr{S}$ to the saliency map:

$$\mathscr{S} = \frac{1}{3}\left(\mathcal{N}(\overline{\mathscr{I}}) + \mathcal{N}(\overline{\mathscr{C}}) + \mathcal{N}(\overline{\mathscr{O}}) + \mathcal{N}(\overline{\mathscr{F}}) + \mathcal{N}(\overline{\mathscr{R}})\right) \tag{14}$$

At any given time, the maximum of the saliency map corresponds to the most salient stimulus, to which attention should be directed next. This maximum is selected by a winner-take-all neural network, [12] which is a biological implementation of a maximum detector.

In the absence of any further control mechanism, the system described so far would direct its attention, in the case of a static scene, constantly to one location, since the same winner would always be selected. To avoid this undesirable behavior Itti *et al.* establish an inhibitory feedback from the winner-take-all (WTA) array to the saliency map (inhibition-of-return;[53]). For our purposes in evaluating dynamic scenes, however, constantly attending to the most salient location in the incoming video is appropriate (but see Discussion for possible improvements), especially since there is little evidence for inhibition-of-return across eye movements in either humans or monkeys. [54] Thus, inhibition-of-return is turned off in our model.

In this work, we only make a very preliminary first pass at integrating task contingencies into this otherwise task-independent model. To this end, we follow the approach recently proposed by Navalpakkam and Itti [55] and add a "task-relevance map" (TRM) between the saliency map and winner-take-all. The TRM acts as a simple pointwise multiplicative filter, making the model more likely to attend to regions of the visual field that have higher task relevance. The product between the saliency and task-relevance maps yields the "attention guidance map" that drives the WTA. While Navalpakkam and Itti have proposed a conceptual framework for populating the TRM based on sequential attention and recognition of objects and their interrelationships, here we set as fairly simplistic task to our model to give preference to image locations that are changing over time. Thus, the TRM in our model is just the inverse difference between the current frame and a sliding average of previous frames.

4. GENERATION OF EYE AND HEAD MOVEMENTS

The covert attention shifts generated by the bottom-up attention model provide inputs to the eye/head movement controller. Since covert attention may shift much more rapidly than the eyes, some spatiotemporal filtering is applied to the sequence of covert shifts, to decide on the next saccade target. In our model, a saccade is elicited if the last 4 covert shifts have been within approximately 10° of each other and at least 7.5° away from current overt fixation.

4.1. Gaze Decomposition

When the head is unrestrained, large amplitude gaze shifts involve coordinated movements of the eyes and head. [24,25] Interestingly, the total displacements of the eyes and head during free-head gaze shifts is well summarized by simple mathematical relationships. It is thus possible to fairly accurately predict eye/head contributions to a given gaze shift, if the initial eye position and desired gaze shift are known. [26,56] The effect of initial eye position onto head/eye movement depends upon the direction of the subsequent eye movement.

For gaze shifts G smaller than a threshold value T, head displacement H is zero. T is given by, with IEP denoting the initial eye position relative to the head and all angular displacements in degrees (the sign of IEP is positive if the eyes are initially deviated in the direction of the subsequent movement, negative otherwise):

$$H = 0 \text{ if } -T < G < T \text{ with } T = \left(\frac{-IEP}{2} + 20\right) \times 0.56 \tag{15}$$

For gaze amplitudes outside this zone, total head movement amplitude H and gaze shift are linearly related such that:

$$H = (1-k) \times T + k \times G \text{ with } k = \left(\frac{IEP}{35} + 1.1\right) \times 0.65 \tag{16}$$

Note that this computation is separately carried out for the horizontal and vertical components of the gaze shifts, rather than for the full two-dimensional shift.

4.2. Saccade dynamics

The dynamics of combined eye-head shifts have been studied extensively. [57–59] The maximum head and eye velocities (Hv and Ev below) are a function of the head and eye movement amplitudes and are given by [60] [p. 551, Fig. 8]:

$$Ev = 473(1 - e^{-E/7.8})^\circ/s \text{ and } Hv = 13H + 16^\circ/s \tag{17}$$

In our implementation, a quadratic velocity profile is computed given the maximum velocity computed above.

In the primate brain, visual input is inhibited during saccades, probably to prevent perception of large motion transients as the eyes move. [61] In our model, the best place to implement such saccadic suppression is the saliency map; thus, during saccades, the saliency map is entirely inhibited. This has three effects: attention is prevented from shifting since there is no more activity in the saliency map; it will take on the order of 200ms for the saliency map to recharge, thus enforcing some intersaccadic latency; and all memory of previous salient visual stimuli will be lost.

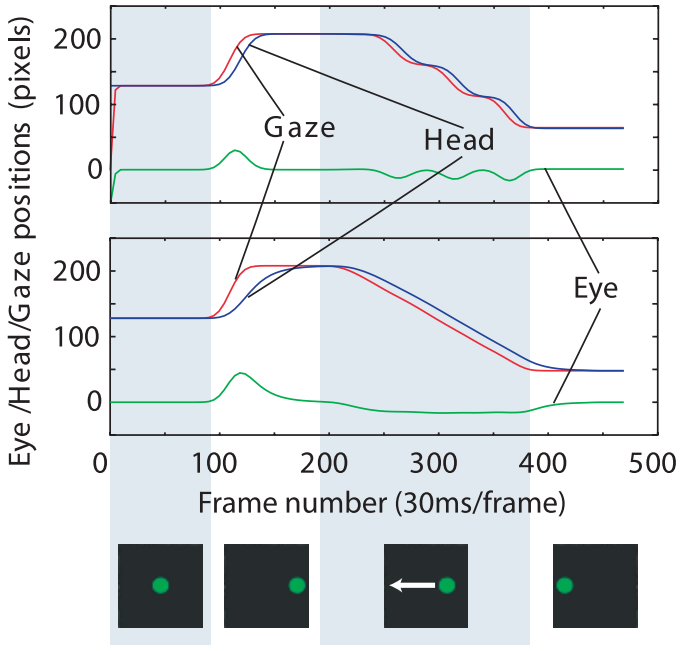

Figure 2. Demonstration of 1D eye/head movement dynamics using a simple synthetic input clip (top: saccade-only model, middle: saccade and smooth pursuit model; bottom: input frames; eye positions are relative to the head). A green disc appears at the center (yielding steady fixation), abruptly jumps rightwards (yielding a rightward eye/head saccade), drifts leftwards (yielding a series of saccades or a smooth pursuit), and finally stops (yielding a final fixation period).

4.3. Smooth Pursuit

The model of Freedman does not include another mode of eye movements that exists in humans and few other animals including in non-human primates, by which the eyes can accurately track a slowly moving target using much slower, non-saccadic, smooth eye and head movements.[7] A very crude approximation of this so-called "smooth pursuit" mode was added to our model, using two simple mass/spring physical models (one for the head and one for the eye): If a gaze shift is too small to trigger a saccade, it will instead become the new anchor point of a spring of length zero linked on its other end to to the eye (or head). The eye/head will thus be attracted towards that new anchor point, and the addition of fluid friction ensures smooth motion. In our implementation, the head spring is five times weaker than the eye spring, and the head friction coefficient ten times stronger than the eye's, ensuring slower head motion **(Figure 2)**.

4.4. Eye Blinks

A final addition to our model is a crude heuristic eyeblink behavior. A blink is initiated during a given simulation time step (0.1ms, i.e., 10,000 fps) if at least 1s has passed since the last blink or saccade, and any of the following:

- with a uniform probability of 10% at every time step if it has been 3s or more since the last blink or saccade, or,
- with a probability of 0.1% if it has been 2s or more, or,
- with a probability of 0.01% if the last 3 attended locations have been close to each other (i.e., stable object), or,
- with a probability of 0.005%.

Blinks last for a fixed 150ms, during which only the winner-take-all (WTA) is fully inhibited so that covert attention is prevented from shifting. Thus, it is assumed here that blinks yield a shorter, weaker inhibition than saccadic suppression. Indeed, the WTA will recover very rapidly from inhibition (10ms or less) so that a blink occurring during smooth pursuit will only minimally disturb pursuit. In contrast, testing with inhibiting the saliency map as was done for saccadic suppression disrupted tracking (pursuit paused after the blink while the saliency map recharged, and a catching-up saccade occurred once covert attention shifts started again). Casual experiments in our laboratory suggest that blinks do not disrupt tracking behavior in humans.

5. ANIMATION AND RENDERING

In order to convincingly illustrate our eye motion synthesis technique, we map the predicted eye/head movements onto a realistic animatable face model. The face model used in our experiments is a digital replica of an actor's face. We used the technique developed by Pighin *et al.*[62] to estimate the geometry of an actor's face from three images. We also extracted a texture map from the images. The eyes were modeled as perfect spheres and textured using the photographs.

The model is controlled through four degree of freedoms: two for the eyes and two for the head. The head's orientation can be changed according to two Euler angles. These two angles control the elevation and azimut of the whole head. The rotation center was estimated from facial motion capture data recorded on the same actor. A head band used as a motion stabilizer provided the rigid motion of the face. We averaged the measured position of the center of rotation over time to provide an estimate of the model's center of rotation. The eyes have similar controls, and rotate around their geometric center independently of the head's orientation.

We convert screen coordinates into Euler angles by assuming that the processed images correspond to a field of view of 90°. Also the full head is considered the same size as the screen and the head's center if rotation is at the same altitude as the center of the screen. Since we do not know the depth of the object in the videos, we assume that the eyes are focused at infinity and point in the same direction.

To make our face behave in a more natural way, we built a mechanism to animate the eyebrows as a function of the direction of the eyes. This mechanism relies on the blending of two shapes. The first one has the eyebrows level, while the second one has raised eyebrows. During the animation we blend these two shapes according to the orientation of the eyes. Upward pointing eyes activate the raised eyebrow blend shape according to a linear ramp **(Figure 3)**.

Figure 3. Sample animation frames.

6. RESULTS

We have applied our model to an unconstrained variety of video segments, including artifical stimuli (e.g., a green disc jumping and drifting across the field of view; **Figures 2, 4**), outdoors video (e.g., a soccer game; **Figure 5)**), and outputs of gaming consoles. Overall, the model attended to locations that made sense to human observers. For example, the system reasonably closely followed the ball, players and overall action in the soccer game, and locked well onto the main character and its ennemies in the console games.

An example of 1D eye/head movement trace is shown in **Figure 2**, demonstrating the three basic modes of operation of the model: steady fixation, saccade, and smooth pursuit. This example shows the shared contribution of head and eye movements during saccades, as expected from Freedman's model, as well as during smooth pursuit. A 2D example is shown in **Figure 4** where a disc is jumping and drifting to various places across the field of view.

There are several failure modes of the model in its present form, typically due to its lack of object recognition capability and of understanding of the contents of the scene. Because we have disabled inhibition-of-return to allow for extended tracking of moving targets, the model currently always attends to the single most salient location in the scene, and ignores all others. This may yield extended periods during which the system tracks a very salient object that a human would consider fairly irrelevant to the overall understanding of the scene and its main action. For instance, in some of the gaming console clips, the screen was populated with a number of indicators of health that were very salient due to their bright

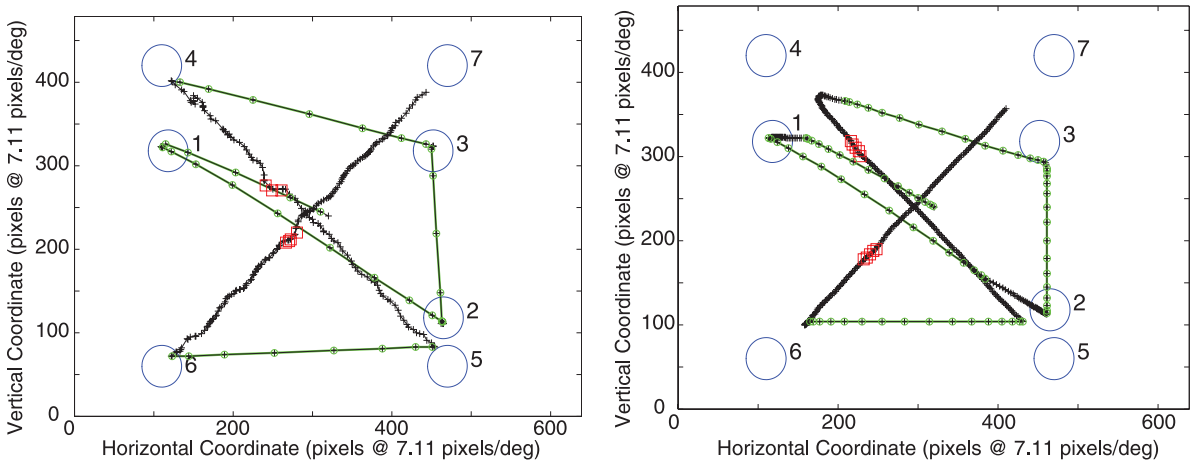

Figure 4. Demonstration of 2D eye/head movement dynamics using a simple synthetic input clip (left: 2D eye trace; right: 2D head trace; crosses are sampled along the traces every 30ms). Eyes and head start at the center of the 640×480 frame (90° horizontal field of view). A disc appears at position '1', and after some delay abruptly jumps to position '2', then '3', then '4', at which point is slowly drifts to '5', jumps to '6' and finally drifts to '7'. Note how the model correctly detects the discs and initiates either a saccade (thicker green traces) or a smooth pursuit. Because the model does not attempt to estimate and predict object trajectories, the eyes and head lie somewhat behind the drifting motion of the target (e.g., does not quite reach targets '5' or '7'). Note how the head motion only accounts for a fraction of the gaze shift (i.e., does not reach all targets), as suggested by Freedman's data. Two eye blinks occured during this simulation (larger square markers on the eye and head traces), which minimally disturbed the tracking behavior.

colors. When these were also animated, they defeated our attempt at focusing the model more onto dynamic than static objects using the task-relevance map. This suggests that the simplistic construction used to populate the task-relevance map, based on the difference between current and previous frames, is inadequate at suppressing very salient animated objects that may be largely irrelevant to the overall action. In the following section we discuss some possible approaches to solving this problem.

Nevertheless, our experimental results reinforce the idea that the overall approach used here indeed is applicable to an unconstrained variety of stimuli, as opposed to more traditional computer vision approaches, which typically are designed to solve a specific task in a specific constrained environment. Indeed, no parameter tuning nor any prior knowledge related to the contents of the video clips was used in any of our simulations, and the exact same model processed the outdoors scenes, games and artificial stimuli. This result is in line with the previously published results by Itti *et al.*, including the reproduction by the model of human behavior in visual search tasks (e.g., pop-out versus conjunctive search); a demonstration of strong robustness to image noise; the automatic detection of traffic signs and other salient objects in natural environments filmed by a consumer-grade color video camera; the detection of pedestrians in natural scenes; and of military vehicles in overhead imagery.[20]

7. DISCUSSION AND CONCLUDING REMARKS

Overall, the proposed system has yielded remarkably robust performance on a wide variety of video inputs. We believe that this result was achieved through our careful modeling of the neurobiology of attention, instead of attempting to develop a dedicated system for specific environments and targets. Further ongoing testing of the algorithm includes comparison between model and human eye movements on the same scenes. Two previous studies[63,64] have already demonstrated good agreement between the covert attention model of Itti *et al.* and eye movements recorded from naive human observers looking at the same static images. Hopefully, our addition of a realistic eye/head movement control subsystem will extend these results to dynamic scenes.

There are many obvious limitations to our model. First, in its present form, it is entirely retinotopic (i.e., all visual processing is made in the coordinate system of the image) and does not account for the well-documented coordinate transforms in parietal cortex and other brain areas.[65] Thus, information about previously attended targets is lost in our model as saccades occur. Second, the saliency map is built entirely from the outputs of local operators, although their receptive fields may be large when they involve coarse pyramid scales. The addition of a prior globally-computed bias for particular locations based on a rapid recognition of the "gist" of the scene[66] may allow the model to more rapidly orient itself towards relevant parts of the incoming visual input. Finally, our treatment of the influence of task onto attentional deployment is at an embryonic stage, as a more detailed implementation first requires that generic object recognition be reasonably well solved in unconstrained scenes.[55]

Nevertheless, the framework presented here enjoys many applications beyond animation of avatars. For instance, foveated video clips are typically three to four times smaller in size than the originals (using MPEG-1 compression), suggesting a potential usefulness of this system in video indexing and archiving. In addition, real-time operation of the attention model has already been demonstrated on a 16-CPU Beowulf computer cluster[67] as well as on a small 4-CPU mobile robot. This suggests potential applications to human/computer interaction and robotics.

A strong limitation of our approach is its computational cost, as it currently cannot be simulated in real-time on a single CPU. Its applicability is thus limited to offline motion generation unless a computer cluster is available. It is thus comparable to complex physical models of human motions, in the sense that it does model accurately low-level mechanisms of the human body at the price of complex and time-consuming simulations. We hope that introducing this model to the graphics community will be beneficial and spur a broader interest in realistic gaze animation. We are studying how to modify this model for real-time performance without dramatically degrading the quality of the motions. One solution that seems particularly promising is to leverage information from the CG scene to detect salient features with little or no image processing.

This work is truly a multidisciplinary effort to merge research results from two communities: computer graphics and neuroscience. We believe that a fertile synergy between the two fields will result in more accurate and realistic models for computer graphics but also will provide validations for some of the theory on low-level human behavior. With this work we not only build a way to realistically synthesize gaze motions but also demonstrate that this model yields a visually plausible representation of low-level attention.

Figure 5. Demonstration of the system in operation on an outdoors video clip (640 × 480 at 30 frames/s, analog interlaced source) processed in a fully automatic manner. Grey arrows indicate current position of covert attention, black arrows current eye position, and white arrows current head position. (1) In this first frame, covert attention, eyes and head are focused onto the soccer ball. (2) The player at the center of the image (orange shirt) starts running towards the ball and becomes highly salient; this has elicited an eye saccade that is just reaching completion in this frame. The head is also moving towards this player, but more slowly so that it lags somewhat behind. (3) Attention, eyes and head are now focused onto that central player and smoothly pursue her. (4) A new player (white shirt) very rapidly enters the field of view and has attracted covert attention, but not in a sufficiently reliable manner to elicit a saccade. (5) Attention, eyes and head smoothly track (with some small lag) the central player (orange shirt), who now has the ball. (6) That player makes a sharp turn and prepares to kick the ball. (7) The player with the white shirt also makes a sharp turn, and is reliably attended to, yieldding a saccade, here caught with the eyes in mid-flight between old (close to head) and new (close to covert attention) positions. (8) Attention, eyes and head regroup close to the player with the white shirt, who is preparing to intercept the ball. No parameter tuning was used to process this or other clips. The visual input is very noisy, low-quality, richly textured and cluttered, and the camera moves substantially, creating large full-field motion transients and changes in object appearances.

Acknowledgements

This research is supported by the National Science Foundation (NSF), the National Eye Institute (NEI), the National Imagery and Mapping Agency (NIMA), and the Zumberge Innovation Research Fund. This paper was in part developed with funds of the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

REFERENCES

1. C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *SIGGRAPH 97 Conference Proceedings*, pp. 353–360, ACM SIGGRAPH, Aug. 1997.
2. M. Brand, "Voice puppetry," in *Proceedings of ACM SIGGRAPH 1999*, pp. 21–28, ACM Press/Addison-Wesley Publishing Co., 1999.
3. T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of ACM SIGGRPAH 2002*, pp. 388–398, ACM Press, 2002.
4. C. Gale and A. F. Monk, "Where am i looking? the accuracy of video-mediated gaze awareness," *Percept Psychophys* **62**(3), pp. 586–595, 2000.
5. J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "Human neural systems for face recognition and social communication," *Biol Psychiatry* **51**(1), pp. 59–67, 2002.
6. J. M. Findlay and R. Walker, "A model of saccade generation based on parallel processing and competitive inhibition," *Behav Brain Sci* **22**, pp. 661–74; discussion 674–721, Aug 1999.
7. D. L. Sparks, "The brainstem control of saccadic eye movements," *Nat Rev Neurosci* **3**(12), pp. 952–964, 2002.
8. L. Spillmann and J. S. Werner, *Visual Perception: The Neurophysiological Foundations*, Academic Press, San Diego, CA, 1990.
9. B. Wandell, *Foundations of vision*, Sinauer Associates, Sunderland, MA, 1995.
10. E. Weichselgartner and G. Sperling, "Dynamics of automatic and controlled visual attention," *Science* **238**(4828), pp. 778–780, 1987.
11. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit Psychol* **12**(1), pp. 97–136, 1980.
12. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol* **4**(4), pp. 219–27, 1985.
13. L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience* **2**, pp. 194–203, Mar 2001.
14. L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of visual behavior*, D. G. Ingle, M. A. A. Goodale, and R. J. W. Mansfield, eds., pp. 549–586, MIT Press, Cambridge, MA, 1982.
15. F. Crick and C. Koch, "Constraints on cortical and thalamic projections: the no-strong-loops hypothesis," *Nature* **391**(6664), pp. 245–50, 1998.
16. J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science* **229**(4715), pp. 782–4, 1985.
17. S. Treue and J. C. Martinez Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature* **399**(6736), pp. 575–579, 1999.
18. M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nat Rev Neurosci* **3**(3), pp. 201–215, 2002.
19. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, pp. 1254–1259, Nov 1998.
20. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research* **40**, pp. 1489–1506, May 2000.
21. J. M. Wolfe, "Visual search in continuous, naturalistic stimuli," *Vision Res* **34**(9), pp. 1187–95, 1994.
22. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modeling visual-attention via selective tuning," *Artificial Intelligence* **78**(1-2), pp. 507–45, 1995.
23. R. Milanese, S. Gil, and T. Pun, "Attentive mechanisms for dynamic and static scene analysis," *Optical Engineering* **34**(8), pp. 2428–2434, 1995.

24. E. Bizzi, R. E. Kalil, and P. Morasso, "Two modes of active eye-head coordination in monkeys," *Brain Res* **40**(1), pp. 45–48, 1972.
25. G. R. Barnes, "Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets," *J Physiol* **287**, pp. 127–147, 1979.
26. E. G. Freedman and D. L. Sparks, "Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: evidence for a gaze displacement command," *J Neurophysiol* **78**(3), pp. 1669–1690, 1997.
27. E. Y. Chen, L. C. Lam, R. Y. Chen, and D. G. Nguyen, "Blink rate, neurocognitive impairments, and symptoms in schizophrenia," *Biol Psychiatry* **40**(7), pp. 597–603, 1996.
28. K. Fukuda, "Eye blinks: new indices for the detection of deception," *Int J Psychophysiol* **40**(3), pp. 239–245, 2001.
29. S. P. Lee, J. B. Badler, and N. I. Badler, "Eyes alive," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 637–644, ACM Press, 2002.
30. T. Rikert and M. Jones, "Gaze estimation using morphable models," in *Int. Conference on Automatic Face and Gesture Recognition*, 1998.
31. R. Stiefelhagen, J. Yang, and A. Waibel, "Estimating focus of attention based on gaze and sound," in *Workshop on Perceptive User Interfaces (PUI '01). Orlando, Florida*, 2001.
32. J. Heinzmann and A. Zelinsky, "3-D facial pose and gaze point estimation using a robust real-time tracking paradigm," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 142–147, 1998.
33. O. Renault, D. Thalmann, and N. M. Thalmann, "A vision-based approach to behavioural animation," *Visualization and Computer Animation* **1**, pp. 18–21, 1990.
34. R. W. Hill, "Perceptual attention in virtual humans: towards ralistic and believable gaze behaviours," in *Simulating Human Agents*, 2000.
35. C. Peters and C. O'Sullivan, "Synthetic vision and memory for autonomous virtual humans," *Computer Graphics Forum* **21**(4), pp. 1–10, 2002.
36. D. Terzopoulos and T. F. Rabie, "Animat vision: Active vision in artificial animals," *Videre: Journal of Computer Vision Research* **1**(1), pp. 2–19, 1997.
37. P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans on Communications* **31**, pp. 532–540, 1983.
38. G. Borgefors, "Distance transformations in digital images," in *CVGIP: Image Understanding*, **54**(2), p. 301, 1991.
39. A. G. Leventhal, *The Neural Basis of Visual Function (Vision and Visual Dysfunction Vol. 4)*, CRC Press, Boca Raton, FL, 1991.
40. H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Seattle, WA*, pp. 222–228, 1994.
41. C. Yee and D. Walther, "Motion detection for bottom-up visual attention," tech. rep., SURF/CNS, California Institute of Technology, 2002.
42. W. Reichardt, "Evaluation of optical motion information by movement detectors," *J Comp Physiol [A]* **161**(4), pp. 533–547, 1987.
43. A. Luschow and H. C. Nothdurft, "Pop-out of orientation but no pop-out of motion at isoluminance," *Vision Research* **33**(1), pp. 91–104, 1993.
44. R. L. DeValois, D. G. Albrecht, and L. G. Thorell, "Spatial-frequency selectivity of cells in macaque visual cortex," *Vision Research* **22**, pp. 545–559, 1982.
45. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J Physiol (London)* **160**, pp. 106–54, 1962.
46. K. S. Rockland and J. S. Lund, "Intrinsic laminar lattice connections in primate visual cortex," *J Comp Neurol* **216**(3), pp. 303–18, 1983.
47. C. D. Gilbert and T. N. Wiesel, "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex," *J Neurosci* **9**(7), pp. 2432–42, 1989.
48. A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, "Visual cortical mechanisms detecting focal orientation discontinuities.," *Nature* **378**(6556), pp. 492–6, 1995.
49. J. B. Levitt and J. S. Lund, "Contrast dependence of contextual effects in primate visual cortex.," *Nature* **387**(6628), pp. 73–6, 1997.

50. M. Weliky, K. Kandler, D. Fitzpatrick, and L. C. Katz, "Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns.," *Neuron* **15**(3), pp. 541–52, 1995.
51. U. Polat and D. Sagi, "The architecture of perceptual spatial interactions.," *Vision Res* **34**(1), pp. 73–8, 1994.
52. M. W. Cannon and S. C. Fullenkamp, "Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations," *Vision Res* **31**(11), pp. 1985–98, 1991.
53. M. I. Posner, Y. Cohen, and R. D. Rafal, "Neural systems control of spatial orienting," *Philos Trans R Soc Lond B Biol Sci* **298**(1089), pp. 187–98, 1982.
54. B. C. Motter and E. J. Belky, "The guidance of eye movements during active visual search," *Vision Res* **38**(12), pp. 1805–15, 1998.
55. V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," in *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany*, pp. 453–461, Nov 2002.
56. E. G. Freedman, "Interactions between eye and head control signals can account for movement kinematics," *Biol Cybern* **84**(6), pp. 453–462, 2001.
57. E. Bizzi, "The coordination of eye-head movements," *Sci Am* **231**(4), pp. 100–106, 1974.
58. J. Lanman, E. Bizzi, and J. Allum, "The coordination of eye and head movement during smooth pursuit," *Brain Res* **153**(1), pp. 39–53, 1978.
59. J. A. Waterston and G. R. Barnes, "Visual-vestibular interaction during head-free pursuit of pseudorandom target motion in man," *J Vestib Res* **2**(1), pp. 71–88, 1992.
60. H. H. Goossens and A. J. V. Opstal, "Human eye-head coordination in two dimensions under different sensorimotor conditions," *Exp Brain Res* **114**(3), pp. 542–560, 1997.
61. A. Thiele, P. Henning, M. Kubischik, and K. P. Hoffmann, "Neural mechanisms of saccadic suppression," *Science* **295**(5564), pp. 2460–2462, 2002.
62. F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin, "Synthesizing realistic facial expressions from photographs.," in *SIGGRAPH 98 Conference Proceedings*, pp. 75–84, ACM SIGGRAPH, July 1998.
63. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Res* **42**(1), pp. 107–123, 2002.
64. R. J. Peters, L. Itti, and C. Koch, "Eye movements are influenced by short-range interactions among orientation channels," in *Proc.Society for Neuroscience Annual Meeting (SFN'02)*, p. 715.12, Nov 2002.
65. P. F. Dominey and M. A. Arbib, "A cortico-subcortical model for generation of spatially accurate sequential saccades," *Cereb Cortex* **2**(2), pp. 153–175, 1992.
66. A. Torralba, "Contextual modulation of target saliency," in *Advances in Neural Information Processing Systems, Vol. 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, Cambridge, MA, 2002.
67. L. Itti, "Real-time high-performance attention focusing in outdoors color video streams," in *Proc. SPIE Human Vision and Electronic Imaging VII (HVEI'02), San Jose, CA*, B. Rogowitz and T. N. Pappas, eds., pp. 235–243, Jan 2002.